Re-Interpreting Mitogenomes

Skujina, Ilze; McMahon, Robert; Hegarty, Matthew

Published in:
Insights in Genetics and Genomics

Publication date:
2017

Citation for published version (APA):
Skujina, I., McMahon, R., & Hegarty, M. (2017). Re-Interpreting Mitogenomes: Are Nuclear/Mitochondrial Sequence Duplications Correctly Characterised in Published Sequence Databases? . *Insights in Genetics and Genomics*, *1*, [6.1].

Insights in Genetics and Genomics

Editorial Open Access

Re-Interpreting Mitogenomes: Are Nuclear/Mitochondrial Sequence Duplications Correctly Characterised in Published Sequence Databases?

Ilze Skujina[1], Rob McMahon[2] and Matthew Hegarty[1]*

[1]Aberystwyth University, IBERS, Gogerddan Campus, UK
[2]Molecular Haematology, Haematology Laboratory, Level 2, Royal Infirmary of Edinburgh, Scotland

*Corresponding author: Matthew Hegarty, Aberystwyth University, IBERS, Gogerddan Campus, Aberystwyth, Ceredigion, SY23 3EE, UK, Tel: +44 (0) 1970 622284; Email: ayh@aber.ac.uk

Original Submission

Received: August 03, 2017
Accepted: August 10, 2017
Published: August 14, 2017

Open Peer Review Status: Editorials, news items, analysis articles, and features do not undergo external peer review.

Acknowledgement: This research was partly funded by the Biotechnology and Biological Sciences Research Council (BBSRC).

How to cite this article: Ilze Skujina, Rob McMahon, Matthew Hegarty. Re-Interpreting Mitogenomes: Are Nuclear/Mitochondrial Sequence Duplications Correctly Characterised in Published Sequence Databases? Insights Genet Genomics. (2017) 1: 6.1

From the late 1970s when Fred Sanger and Alan Coulson developed their chain termination method for rapid determination of DNA sequence [1], Sanger sequencing dominated nucleic acid and genome research until about a decade ago when technical developments led to the "next big thing" in sequencing genomes. In 2005, the first workable sequencing strategies entailing arrays of millions of DNA templates sequenced in parallel became publicly available [2,3]. It was not long before such high-throughput, massively-parallel approaches transformed biological research, generating vast amount of genome data. However, for the successful interpretation of sequence data, the quality of primary sequence reads and correct assembly of contigs is paramount. While no sequencing technique is error-free, and biases have been accounted for with both first generation and next generation sequencing (NGS) technologies, when it comes to "difficult genomic regions", it can be anticipated that results relying solely on just one sequencing approach may produce fundamental errors within published sequences, which are difficult if not impossible to identify without closer examination.

Mitochondria are the major energy providers in eukaryotes. They are cytoplasmic, semi-autonomous organelles with the majority of their ~16.5kbp (in a standard vertebrate) original precursor genome now integrated into the nuclear genome of the cell but with a small portion consisting of 37 genes and ~1kb of non-coding sequence referred to as the control region (CR) remaining within the organelle itself [4]. Due to their high abundance, small genome size relative to nuclear DNA (nuDNA), and non-recombinant inheritance, mitochondria have been extensively employed in genetic studies. For example, mitochondrial DNA (mtDNA) has been widely used in forensic investigations [5], archaeogenomic research [6,7], population genetics to establish genetic relationships, as well as molecular systematics and reconstruction of species history [8,9]. Equally, the quality and quantity of mtDNA is often employed as a marker of mitochondrial activity and, considering its bioenergetics role within cells, mtDNA variants and defects have been implicated in a plethora of pathologies, metabolic syndromes (such as diabetes), aging, aging-associated degenerative diseases and cancer [10,11]. Consequently, if it is the mtDNA sequence on which a whole populations' history is based, aetiology of a disease explained, or on which a judicial verdict depends – the sequence assembly must be 100% accurate.

However, even with the current power of NGS platforms and decades of experience with Sanger sequencing, erroneous mitogenome sequences are published, have been deposited, and remain available within the databases. There are three main biological issues that have to be accounted for in order to obtain a true and accurate mitochondrial sequence. 1) mitochondrial heteroplasmy: the high copy number that makes mtDNA an easily accessible and attractive tool for population genetics and ancient DNA studies may prove a tripping stone as sequence genotype can vary from organelle to organelle between different cell/tissue types and even over time [12]. 2) the fragments of mtDNA that are also integrated within germline nuclear sequences – referred to as "numts" (nuclear mitochondrial sequences, [13-16]): while these "molecular fossils" provide exciting opportunities to study mtDNA and species evolution [15,17-18], it is not uncommon that numt sequences are mistaken for authentic mtDNA and included within mitosequences [19,20]. 3) mitochondrial gene duplications/deletions: a phenomenon associated with abnormality in mammals [21], but which have been found in normally functioning mitochondria in other organisms [22-24]. Birds are a particularly good example, as mitochondrial gene duplications and/or non-coding control region duplications (YCR), and different arrangements of the gene order have been observed arisen independently multiple times across the avian family [25-28].

When sequencing avian mitochondrial genomes, all of these three impediments will have to be taken into account: even though numts and heteroplasmy have been shown to exist in a wide number of taxa [16,29-30] and tandem repeat sequences within the CR have been observed in many mammalian species such as horse, deer, shrew, bat and various carnivores [31-35], birds are the only endotherms with reported genic duplications within the actual mitochondrial DNA [26,28,36,37]. To complicate matters further, some birds have been recorded to possess portions of the CR duplications integrated within the nuclear genome and with heteroplasmy in terms of the repeat number within the variable domain of the YCR in mtDNA and/or nuDNA. For instance, a set of consistent underlying peaks at 5-20% of maximum peak intensity can be clearly determined when analysing Sanger sequencing chromatogram of the YCR-specific PCR amplicons of the mitochondria of the Red kite (*Milvus milvus*) [38]. The double signal, which is present only in female birds, originates from the presence of mitochondrial DNA sequences translocated to the W chromosome [39,40]. Illumina shotgun sequencing failed to characterise the mitochondrial and W-chromosome duplications, as most of the reads from these regions remained unmapped because of the highly repetitive and relatively long repeat structure (~1.5kb), thus they were discarded by the assembly package in the first place, during early contig assembly. A recent study by Nacer and do Amaral reported a striking pseudogenization in avian phylogenetics and concluded that avian numts may be much more frequent and longer than previously thought [41]. For instance, a nuclear copy of mtDNA covering 93.6% of the mitogenome was found in the Peregrine falcon and numt sequences in falcons totalled ~49kb or ~0.004% of the whole nuclear genome. In another instance, a phylogenetic study of cranes based on standard PCR amplification of the mtDNA sequences by Krajewski et al. [42] did not report evidence for a duplication within crane mtDNA. Most of the mtDNA sequences used were obtained from ~500-1000bp overlapping amplicons providing ≥2× coverage for ~25% of the mtDNA molecule. ND6 and CR sequences of all cranes were obtained from Krajewski et al. [43] and Fain [44], respectively, and incorporated within the sequence assembly. For seven years the deposited sequences

were believed to be a complete and accurate mitochondrial picture of the Crane (Gruidae) family members and these sequences are still accessible via NCBI [45]. However in 2017, Akiyama et al. [46] using a mix of initial long-range PCR (LR-PCR) and then nested PCRs to study the structure of mtDNA, showed that all the 13 analysed *Gruidae* species previously sequenced by Krajewsi et al. [42] possessed a duplication block consisting of Cytb, tRNAThr, tRNAPro, ND6, tRNAGlu and CR. The duplication was conserved across all the species and was similar to those detected in other unrelated avian species such as albatrosses (36,47], spoonbills [48], and boobies [49], in a stroke increasing the average size of Gruidae mitochondrial genome from 16.5kb to ~22kb. This error not only changed the perception of the linage divergence of the cranes, but also radically altered our understanding of the evolution of the YCR within the avian phylogeny [37,46,49,50].

To conclude, living in an era when obtaining large quantities of sequencing data is no longer an obstacle, the effort should be focused on ensuring its correct validation. Whilst no sequencing technique is perfect and warning signs have been issued on numerous occasions [51,52], strategic approaches to dealing with "difficult sequence" such as mtDNA duplications, deletions and inerchromosomal transfers should be developed, whether these be mixtures of laboratory-based techniques such as LR-PCR, sequence capture and use of more than one sequencing approach (i.e. combining Sanger with NGS or use of the new long-read sequencers such as MinION) or the generation of novel bioinformatics pipelines for resolving these issues in existing datasets.

References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences. 1977; 74: 5463-5467.
2. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437: 376-380.
3. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005; 309: 1728-1732.
4. Taanman JW. The mitochondrial genome: structure, transcription, translation and replication. Biochimica et Biophysica Acta. 1999; 1410: 103–123.
5. Holland MM, Parsons TJ. Mitochondrial DNA sequence analysis-validation and use for forensic casework. Forensic Science Review. 1999; 11: 21-50.
6. Schurr TG, Ballinger SW, Gan YY, Hodge JA, Merriwether DA, at al. Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages. American Journal of Human Genetics. 1990; 46: 613.
7. Larson G, Karlsson EK, Perri A, Webster MT, Ho SY, et al. Rethinking dog domestication by integrating genetics, archeology, and biogeography. Proceedings of the National Academy of Sciences USA. 2012; 109: 8878-8883.
8. Avise JC. Molecular Markers, Natural History and Evolution. New York: Chapman and Hall. 1994.
9. Liu CZ, Wei GH, Hu JH, Liu XY. Complete mitochondrial genome of the Swan Goose (*Anser cygnoides L.*) and its phylogenetic analysis. Mitochondrial DNA - Part A. 2016; 27: 2427-2428.
10. Wallace DC. Mitochondrial DNA variation in human radiation and disease. Cell. 2015; 163: 33-38.
11. Wallace DC. Genetics: Mitochondrial DNA in evolution and disease. Nature. 2016; 535: 498-500.
12. Wallace DC. Mitochondrial DNA sequence variation in human evolution and disease. Proceedings of the National Academy of Sciences USA. 1994; 91: 8739–8746.
13. Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. Journal of Molecular Evolution. 1994; 39: 174–190.
14. Zischler H, Geisert H, von Haeseler A, Pääbo S. A nuclear ''fossil'' of the mitochondrial D-loop and the origin of modern humans. Nature. 1995; 378: 489–492.
15. Bensasson D, Zhang D-X, Hartl DL, Hewitt GM. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends in Ecology & Evolution. 2001; 16: 314–321.
16. Hazkani-Covo E, Zeller RM, Martin W. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genetics. 2010; 6: 1000834.
17. Kim JH, Antunes A, Luo SJ, Menninger J, Nash WG, et al. Evolutionary analysis of a large mtDNA translocation (numt) into the nuclear genome of the *Panthera* genus species. Gene. 2006; 366: 292-302.
18. Karanth KP. Primate numts and reticulate evolution of capped and golden leaf monkeys (Primates: Colobinae). Journal of Biosciences. 2008; 33: 761-770.
19. Wallace DC, Stugard C, Murdock D, Schurr T, Brown MD. Ancient mtDNA sequences in the human nuclear genome: a potential source of errors in identifying pathogenic mutations. Proceedings of the National

Academy of Sciences. 1997; 94: 14900–14905.

20. Yao YG, Kong QP, Salas A, Bandelt HJ. Pseudomitochondrial genome haunts disease studies. Journal of Medical Genetics. 2008; 45: 769–772.
21. Damas J, Samuels DC, Carneiro J, Amorim A, Pereira F. Mitochondrial DNA rearrangements in health and disease—a comprehensive study. Human Mutation. 2014; 35: 1-14.
22. Moritz C, Brown WM. Tandem duplication of D-loop and ribosomal RNA sequences in lizard mitochondrial DNA. Science. 1986; 233: 1425-1428.
23. Boore JL. The duplication/random loss model for gene rearrangement exemplified by mitochondrial genomes of deuterostome animals. In: Comparative Genomics. The Netherlands: Springer. 2000; 133-147.
24. Lavrov DV, Boore JL, Brown WM. Complete mtDNA sequences of two millipedes suggest a new model for mitochondrial gene rearrangements: duplication and nonrandom loss. Molecular Biology and Evolution. 2002; 19: 163-169.
25. Gibb GC, Kardailsky O, Kimball RT, Braun EL, Penny D. Mitochondrial genomes and avian phylogeny: complex characters and resolvability without explosive radiations. Molecular Biology and Evolution. 2007; 24: 269–280.
26. Haring E, Riesing MJ, Pinsker W, Gamauf A. Evolution of a pseudo-control region in the mitochondrial genome of Palearctic buzzards (genus Buteo). Journal of Zoological Systematics and Evolutionary Research. 1999; 37: 185–194.
27. Mindell DP, Sorenson MD, Dimcheff DE. Multiple independent origins of mitochondrial gene order in birds. Proceedings of the National Academy of Sciences of the United States of America. 1998; 95: 10693-10697.
28. Singh TR, Shneor O, Huchon D. Bird mitochondrial gene order: insight from 3 warbler mitochondrial genomes. Molecular Biology and Evolution. 2008; 25: 475-477.
29. Kumazawa Y, Ota H, Nishida M, Ozawa T. Gene rearrangements in snake mitochondrial genomes: highly concerted evolution of control-region-like sequences duplicated and inserted into a tRNA gene cluster. Molecular Biology and Evolution. 1996; 13: 1242-1254.
30. Kmiec B, Woloszynska M, Janska H. Heteroplasmy as a common state of mitochondrial genetic information in plants and animals. Current Genetics. 2006; 50: 149-159.
31. Xiufeng X, Árnason Ú. The complete mitochondrial DNA sequence of the horse, *Equus caballus*: extensive heteroplasmy of the control region. Gene. 1994; 148: 357-362.
32. Hoelzel AR, Lopez JV, Dover GA, O'Brien SJ. Rapid evolution of a heteroplasmic repetitive sequence in the mitochondrial DNA control region of carnivores. Journal of Molecular Evolution. 1994; 39: 191-199.
33. Fumagalli L, Taberlet P, Favre L, Hausser J. Origin and evolution of homologous repeated sequences in the mitochondrial DNA control region of shrews. Molecular Biology and Evolution. 1996; 13: 31-46.
34. Cook CE, Wang Y, Sensabaugh G. A Mitochondrial Control Region and Cytochrome B Phylogeny of Sika Deer (*Cervus nippon*) and Report of Tandem Repeats in the Control Region. Molecular Phylogenetics and Evolution. 1999; 12: 47-56.
35. Wilkinson GS, Mayer F, Kerth G, Petri B. Evolution of repeated sequence arrays in the D-loop region of bat mitochondrial DNA. Genetics. 1997; 146: 1035-1048.
36. Abbott CL, Double MC, Trueman JWH, Robinson A, Cockburn A. An unusual source of apparent mitochondrial heteroplasmy: duplicate mitochondrial control regions in Thalassarche albatrosses. Molecular Ecology 2005; 14: 3605–3613.
37. Eberhard JR, Wright TF. Rearrangement and evolution of mitochondrial genomes in parrots. Molecular Phylogenetics and Evolution. 2016; 94: 34-46.
38. Skujina. Population genetics of an endangered bird of prey: the Red Kite in Wales. MPhil dissertation, IBERS, Aberystwyth University, Aberystwyth, UK. 2013.
39. May CA, Wetton JH, Parkin DT. Polymorphic sex-specific sequences in birds of prey. Proceedings of the Royal Society, London, Series B. 1993; 253: 271-276.
40. Skujina I, McMahon R, May CA, Wetton JH, Hayward M, et al. High levels of mitochondrial DNA heteroplasmy and mitochondrial pseudo-control region transposition onto the sex-chromosome in the Red Kite (*Milvus milvus*). In prep.
41. Nacer DF, do Amaral FR. Striking pseudogenization in avian phylogenetics: numts are large and common in falcons. Molecular Phylogenetics and Evolution. 2017; 115: 1-6.
42. Krajewski C, Sipiorski JT, Anderson FE. Complete mitochondrial genome sequences and the phylogeny of cranes (gruiformes: Gruidae). Auk. 2010; 127: 440–452.
43. Krajewski C, Fain MMGG, Buckley L, King DG. Dynamically heterogenous partitions and phylogenetic

inference: an evaluation of analytical strategies with cytochrome b and ND6 gene sequences in cranes. Molecular Phylogenetics and Evolution. 1999; 13: 302–313.

44. Fain MG. Phylogeny and evolution of cranes (Aves: Gruidae) inferred from DNA sequences of multiple genes. Ph.D. dissertation, Southern Illinois University, Carbondale. 2001.

45. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Research. 2014; 42: D32-37.

46. Akiyama T, Nishida C, Momose K, Onuma M, Takami K, et al. Gene duplication and concerted evolution of mitochondrial DNA in crane species. Molecular Phylogenetics and Evolution. 2017; 106: 158-163.

47. Eda M, Kuro-o M, Higuchi H, Hasegawa H, Koike H. Mosaic gene conversion after a tandem duplication of mtDNA sequence in Diomedeidae (albatrosses). Genes & Genetic Systems. 2010; 85: 129–139.

48. Cho H-J, Eda M, Nishida S, Yasukochi Y, Chong J-R, et al. Tandem duplication of mitochondrial DNA in the black-faced spoonbill, *Platalea minor*. Genes & Genetic Systems. 2009; 84: 297–305.

49. Morris-Pocock JA, Taylor SA, Birt TP, Friesen VL. Concerted evolution of duplicated mitochondrial control regions in three related seabird species. BMC Evolutionary Biology 2010; 10: 14.

50. Zhou X, Lin Q, Fang, Chen X. The complete mitochondrial genomes of sixteen ardeid birds revealing the evolutionary process of the gene rearrangements. BMC Genomics. 2014; 15: 573.

51. Shi NN, Fan L, Yao YG, Peng MS, Zhang YP. Mitochondrial genomes of domestic animals need scrutiny. Molecular Ecology. 2014; 23: 5393–5397.

52. Peng, M-S, Shi N-N, Yao Y-G, Zhang Y-P. Caveats about interpretation of ancient chicken mtDNAs from northern China. Proceedings of the National Academy of Sciences. 2015; 112: E1970-1971.